Tijdschrift voor Economie en Management
Vol. L, 4, 2005

New Trends in Data Mining

by J. HUYSMANS, B. BAESENS, D. MARTENS,
K. DENYS and J. VANTHIENEN

Johan Huysmans
KULeuven, Departement Toegepaste Economische Wetenschappen, Leuven

Bart Baesens
KULeuven, Departement Toegepaste Economische Wetenschappen, Leuven

David Martens
KULeuven, Departement Toegepaste Economische Wetenschappen, Leuven

Katrien Denys
KULeuven, dienst Oftalmolotie, Leuven

Jan Vanthienen
KULeuven, Departement Toegepaste Economische Wetenschappen, Leuven

ABSTRACT

The amount of newly created information increases every year. Large-scale automation projects, the ubiquity of personal computers and the declining prices of storage are all factors that contribute to this trend. The huge amount of information has made it impossible for human analysts to gain a deeper understanding of their data without at least some form of computer-aid. Data mining can be used to automate this process of knowledge discovery from databases.
Over the past years, data mining has grown from a relatively unknown technique into a widespread billion dollar business. While data mining was first only adopted in the retail and banking sectors, we can nowadays observe a proliferation of the application domains. In this paper we cover some of these recent application domains and explain how data mining can contribute towards providing new insights and an increased efficiency in these fields. In the second part of this article, we present some new data mining techniques that are expected to make a rapid transition into business environments.

I. INTRODUCTION

Data Mining is formally defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad (1996)). From a relatively unknown technique, adopted by some credit institutions and retailers, data mining has grown into a billion dollar business. Banks use data mining to determine the credit worthiness of their applicants, retailers adopt it to choose the optimal layout for their shops and insurance companies rely on data mining to identify potential fraudulent claims.

That data mining has grown to maturity can also be observed from the fact that most database suppliers offer integrated data mining solutions. These tools facilitate the process of knowledge creation and contribute towards the further spread of data mining. In some domains, like crime prevention or bio-informatics, data mining is still in its infancy. An unseen optimism characterizes most of these new applications and the known restrictions of data mining are easily forgotten. Besides the proliferation of the potential application domains, it can also be observed that improvements to the existing techniques are continuously being made. These improvements take place in many different areas: more precise and comprehensible forecasts, integration with existing databases, real-time analyses, ...

The structure of this article follows this two-part approach. In the first part, we discuss a number of application domains in which data mining is expected to play a crucial role in the near future. In the second part, several new techniques are presented which promise to offer significant improvements over existing techniques.

II. NEW APPLICATIONS OF DATA MINING

A. *The fight against terror: the needle in the haystack*

In the aftermath of the September 11 attacks, many countries approved new laws in the fight against terrorism. These laws allow intelligence services to gather all information deemed necessary to prevent new attacks and to swiftly identify potential terrorists. In this domain, the United States of America played a pioneer role with their *Total Information Awareness* program. The goal of this program was the creation of a huge central database that consolidates all the available information on the population. Similar projects were announced in Europe and the rest of the world. Although some of these programs were cancelled due to massive resistance of privacy organizations, most of these plans nevertheless seem to resurrect later under a slightly

different name. For example, the “Total Information Awareness” program was conveniently relabeled to “Terrorist Information Awareness” program.

Combining information originating from the private sector, like bank and purchase information, with government information produces a potential treasure of information but aside the already mentioned privacy problems, several other challenges arise. A first problem concerns the diversity and heterogeneity of the data. Besides structured information, the central database must be able to deal with text- and multimedia objects. This data diversity provides specific problems for most data mining algorithms that were usually developed to recognize patterns in structured data. Additionally, the amount of data seriously restricts the scalability of the algorithms. The execution time can rise only to a certain extent when the quantity of available data increases. Real-time applications also pose strict limits on the allowed execution time of the algorithms. For example, 230 cameras were placed in the city of London to control the traffic towards the center and to automatically read the license plates of the passing vehicles. With an estimated 40.000 vehicles that pass the cameras every hour, the system must recognize 10 vehicles each second. An immense task, that poses heavy requirements on both hard– and software (Transport for London (2004)).

Finally, one has to take into consideration the costs that are associated with every decision. Systems that provide the most accurate predictions will often be inferior to less accurate systems when taking into account misclassification costs. For example, a system that is able to identify all possible airplane hijackers but misclassifies some normal passengers as terrorists will be preferred above a system that classifies most of the passengers correct but misclassifies some of the terrorists. The inconvenience that is being experienced by normal passengers is of minor importance in comparison to the damage that one terrorist can cause. But the choice among different systems is not an easy one: to develop optimal systems one must be able to quantify all costs and benefits and this might be a difficult task. For example, how do we measure the inconvenience for the passengers? Besides the measurable delays during check-in, many other factors play a role: feelings of (in)security, deterrence of potential terrorists,..

It is clear that many problems must be overcome before the potential of data mining can be fully exploited in this area. To our opinion, the problems that are easiest to tackle are the problems of technical nature. We are convinced that further research in this rapidly evolving domain will provide solutions for most of the current restrictions. The problems that are harder to overcome are the ones of social nature because personal freedom is embraced by most people as one of the most-valued principles. Heated (and healthy) debates will continue to discuss the trade-off that must be considered: how much privacy is one willing to give up for a (difficult-to-measure) increase in safety?

B. *Bioinformatics: the quest for cures*

A second domain in which data mining is strongly embraced is bio-informatics. Bio-informatics is the science concerning management, mining and interpretation of biological sequences and structures. Genome sequencing projects have contributed to an exponential increase in complete and partial sequence databases. The Structural Genome Initiative has the aim of cataloguing the structure-function information of proteins. Progress in technologies such as microarrays, resulted in the beginning of the subdomains of genomics and proteomics. These fields study the genes, proteins and the circuit inside the cell that regulates the gene-expression. Lots of data is being generated, data that must be mined if mankind ever wants to expose the mysteries of cells.

During the last years, enormous progress has been achieved, but there remain a number of fundamental problems in bio-informatics, such as predicting protein structures and finding genes. Data mining will play a fundamental role in the understanding of gene expression, the development of medicines and other problems in genomics and proteomics. Furthermore, text mining will be important to filter knowledge from the growing offer of literature concerning bioinformatics.

C. *Retail: the second wave*

Retailers were always on the front-line for the adoption of data mining. Since longtime association analysis is used to examine which products are frequently sold together and based on this analysis retailers can choose the optimal lay-out of the shelves to maximize profit.

But also in this area, it seems that data mining has hardly grown out of its infancy. Several retailers experiment with self-scanners, devices that allow the customers to scan the bar codes of the products. One advantage is that the shop clerks no longer have to scan all the products themselves, but that they only have to take care of the payment. Sporadically, a check of one's basket is done to avoid fraud. The distributors promote this new shopping method as the ideal way to avoid long queues at the checkout, but more importantly, the self-scanners create lots of interesting information. With traditional shopping, the chains are only able to find out which products were bought by a certain customer. By promoting self-scanning, the retailers can obtain worthwhile extra information: they can follow perfectly how their customers wander through the shop. This provides new possibilities for data mining. Alternatives to the traditional association rules which profit of this additional information have already been presented for this task (Huysmans et al. (2004)).

Self-scanning offers other exciting opportunities. We expect that real-time discounts will make their entry in the coming years. While scanning a product, a discount that is valid for some minutes is offered for a related product. Thus, the distributor provokes impulse purchases and is able to see

which customer is subject to which discount. Privacy organizations stand of course very fearful towards such techniques and that this apprehension is not unfounded becomes clear from the tale of the German retailer Metro Group (CASPIAN (2004)). This distributor placed without knowledge of its customers RFID-tags in their loyalty cards. These tags, small senders who transmit the customer number, could subsequently be used for following the customers through the shop, but possibly also . Under large pressure of various privacy organizations, the chain had to return on its steps and remove the RFID-tags from the loyalty cards. However, we can also expect many advantages from this successor of bar codes. Attaching RFID-tags to cattle enables an efficient control of the food chain. In courier services and stock management centers this technology can lead to considerable efficiency improvements.

In the retail sector, we can also expect the advent of location based services or L-commerce. The goal of L-commerce is to use data about the current position of a person to provide him with targeted and relevant information. For example, while wandering through a shopping street a person is continuously informed about the special promotions of the shops he passes by. But the potential applications go further, especially when L-commerce is intertwined with other techniques like RFID and web access. For example, when passing a clothing shop the RFID-scanners in the display window pick up information from the RFID-tags in the clothes and loyalty card of the passant. Based on this information, a database is inquired to find those clothes that fit together with the clothes the customer is currently wearing or has bought in the past. The selected clothes can then be shown immediately on the mobile phone of the person passing by in combination with the clothes the customer is wearing.

It can still take many years before these stories will become reality. GPS-enabled mobiles and PDA's are still expensive but prices continue to decline every month and the opportunities for marketers are so enormous that we expect the devices to be given away for free in return for advertisements being shown.

D. *Semantic Web: an impossible dream or near future?*

A little more than a decade ago, the brainchild of Tim Berners-Lee saw the daylight and this was the starting point for the Internet to become one of the most important technologies of the past century. Only few inventions have had a larger impact on the way businesses work than the World Wide Web. Nowadays, the Internet has become the instrument de-facto for the dissemination of information of any kind: research results, library access, corporate information, everything is made publicly available by means of web pages and some hyperlinks.

But the impressive growth of the Internet has also lead to some serious drawbacks. The advantage that everyone, even the technologically

ignorant, is able to add pages has also turned into a serious drawback. Links that work today can suddenly stop working because someone changes the directory structure of the site that is being referred to. But the main point of critique is the fact that almost all web pages are only intended to be readable by humans. The mixture of content and lay-out tags into one document makes them unsuitable for machine processing. Based on the title of a web page and words that appear in the document, a search engine tries to grasp the meaning of a web page but this approach often fails. For example, a search for the exact birthday of Albert Einstein on the internet with the keywords 'birthday' and 'Einstein', returns several thousands of results but many of these results are unrelated to our inquiry. Wouldn't it be nice to receive an exact answer on our query without having to scan through all these pages?

To fulfill this goal it is necessary to complement web documents with machine-readable meta-data and that is exactly what the second generation Internet, coined the Semantic Web or SemWeb, promises to realize. The basic technology underlying the semantic web is the Resource Description Framework (RDF) which is used to describe resources. A resource can be anything: a web page, a person, an intangible good and sets of triples are used to describe these resources. Every triple consists of the following: the first part refers to the resource, the second part to an attribute of the resource and the third part provides the value that the resource has for this attribute. We will clear this up with a small example. Let us show how it is possible to indicate that Albert Einstein was born on 14 March 1879 in RDF:

<Albert Einstein><was born><14 March 1879>

However, this notation is a little oversimplified. Normally, there is a Universal Resource Identifier (URI) associated with every resource and attribute. Suppose that the URI of Albert Einstein is 'http://www.persons.com/AlbertEinstein' and the URI of 'was born' corresponds with 'http://xmlns.com/foaf/0.1/DateOfBirth'.
The above RDF-triple can then be rewritten as

<http://www.persons.com/AlbertEinstein><http://xmlns.com/foaf/0.1/DateOfBirth> …<14 March 1879>

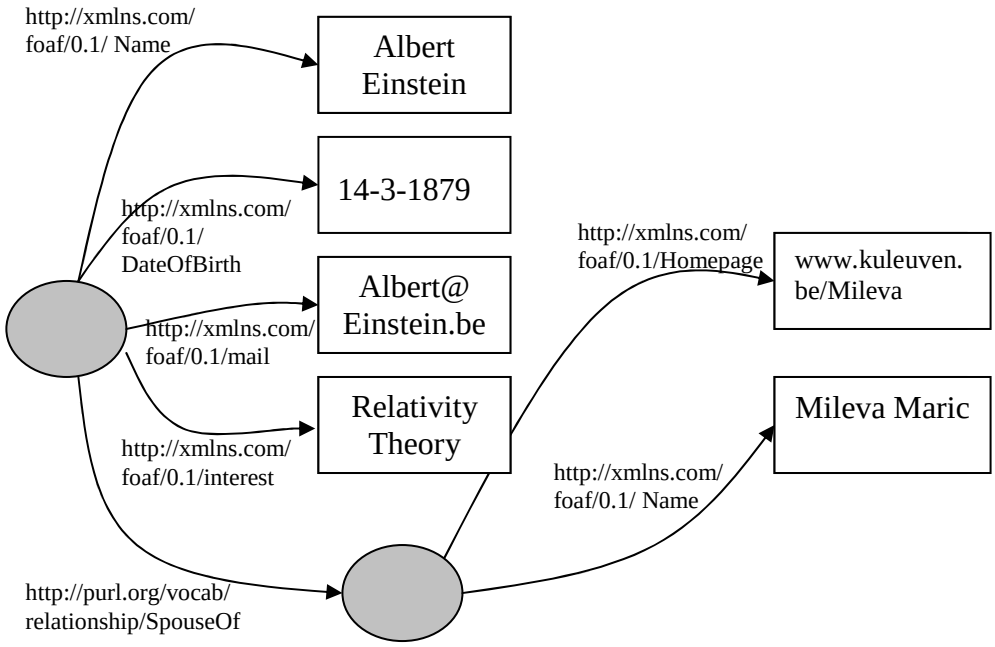
The URI's are chosen rather arbitrarily and it is therefore not necessary that they refer to an actual existing web page. Second, the value-part can be either a literal or another resource. For example, a possible way to indicate that Albert Einstein was married to Mileva Marić is as follows:

<http://www.persons.com/AlbertEinstein><http://purl.org/vocab/relationship/spouseOf><http://www.persons.com/MilevaMaric>

In this statement, the value-part is also a resource that at its turn can be described with the aid of RDF. Other statements can be added to obtain a

detailed machine readable description of Albert Einstein. There are several possible methods to write the above RDF-statements. In Figure 1, the two above statements are shown in a graph-format together with some additional information. From this graph, both a machine and a human can find out when Einstein was born or what his particular topic of interest is.

FIGURE 1
RDF-Description in graph format of Albert Einstein



There is however one problem: the URI's in this example were arbitrarily chosen and it is therefore quite likely that other people will use different URI's to denote the same resources. This is a tremendous problem: suppose that someone provides additional information about Albert Einstein but that he uses a different URI. Human readers will probably be able to find out that both resources refer to the same person, but for computer processing this is rather difficult. Similarly the URI of an attribute can be arbitrarily chosen by the person creating the RDF statements. There is no general solution for this problem. However, several projects have proposed an immense number of standardized URI's. For example, the Friend-of-a-Friend project (FOAF (2005)) has created a significant number of tags that can be used to describe people, the relationships between them and the projects they

are working on. The Dublin Core MetaData Initiative (Dublin Core (2005)) provides standard URI's to indicate who is the creator of a text, what the title and subject is, and so on. But keep in mind that most users will never have to make a decision about which URI's he or she will use. Programs will automate this task and the end user will probably not even be aware that he is creating RDF-statements. Furthermore, automatic creation of RDF-triples from relational databases is relatively straightforward.

The potential of the semantic web is enormous and data mining can both contribute towards and profit from the creation of this new web behind the existing Internet. First, web mining can help in building semantic structures by extracting information from the available documents on the internet. Second, the results of web mining can be improved by exploiting the available semantic structures. It would be possible to merge information from many different sources and instead of performing simple word-matching search engines could drastically improve the relevance of the returned documents or guide the user towards the desired information. For example, in production environments the manufacturer of a product could provide RDF-statements describing the characteristics of a certain product. Customers could then indicate their requirements and an automatic matching with the characteristics could be performed. The web can become one gigantic marketplace where offer and demand can automatically find each other.

There are still many questions to be resolved before this semantic web will be reality. First, as everyone is able to add, alter and remove his RDF-statements, just like normal web pages, who will combine all these RDF-statements to extract meaningful information? The main potential of the semantic web lies in combining statements coming from many different sources. Will gigantic search engines continue to structure all the available information or can we expect that personal search agents will be adopted to find a better match with the user's preferences. Second, as everyone is able to make statements, we will have to deal with the possibility that erroneous RDF-statements are encountered. For example, a restaurant owner may declare wrong opening hours for a competing restaurant. Different levels of confidence might therefore be assigned depending on the origin of the statements.

So, it can be concluded that the semantic web provides both tremendous opportunities and challenges. Data Mining might be ideal to facilitate the construction of this web and could at the same time profit enormously if useful information from many different sources becomes available for machine-processing.

E. *Other applications*

There are many domains other than the four already mentioned above – terror prevention, bio-computing, market basket analysis and semantic web – where data mining plays or will play an important role. For example in astronomy,

where new telescopes provide very detailed pictures of the universe, data mining techniques are used for an automatic classification of newly discovered celestial bodies. In text processing, many text mining projects emerge that create information from unstructured text documents. For example, the `Google News'-site (2005) uses clustering techniques to group news messages from several news providers according to the subject. Automatic summarizing of texts or detecting spam in e-mail messages are other applications in which text mining plays a vital role. The results of search engines depend strongly on the progress of text mining and natural language understanding. Search engines which can answer simple questions are already available and improve continuously, but it will still take many years before a computer can endure a Turing-test (Turing (1950)).

III. NEW TECHNIQUES IN DATA MINING

Not only do the application areas of data mining expand continuously, but also the utilized techniques keep up improving. In the rest of this article we take a closer look at four new methods: multi-relational data mining, support vector machines, Bayesian networks and ensemble methods.

A; *Multi-relational data mining*

Most data mining algorithms are propositional; this means that they were devised to discover patterns in a single data table. However, larger databases generally contain several tables between which a number of relations have been defined. The example database of Table 1 consists of two tables, whereby the second table indicates which persons from the first table are married with each other (Dzeroski (2003)). From this database, one wants to establish a decision tree so that the important customers can be identified swiftly. Propositional algorithms create classification rules of the following form: IF (income > 108000) THEN important customer = YES

Observe that only the information from the first table was used for the creation of this rule. Relational algorithms on the other hand are able use the relationships that exist between the tables.

An example of such a rule is:

IF (x is married with a person with income > 10800) THEN important customer (x) = YES

FIGURE 2
Relational Database with two tables (based on Dzeroski, 2003)

Customer Table

ID	Gender	Age	Income	Expense	Important Customer
C1	Male	30	214000	18800	Yes
C2	Female	19	139000	15100	Yes
C3	Male	55	50000	8600	No
C4	Female	48	26000	8600	No

MarriedWith Tablel

Partner 1	Partner 2
C1	C2
C3	C4

The large advantage of these multi-relational algorithms is that they are immediately applicable to the omnipresent relational databases. There is no preparation step necessary whereby the data is incorporated into one table such as is necessary for the propositional algorithms. A second advantage is that the relational rules are frequently more expressive. A larger number of possible hypotheses can be expressed. However, this provides also the necessary problems: by the large number of hypotheses that one can form, the probability that one of these hypotheses is accidentally significant for the given data increases too. For example: *blonde women of age 45 married with a small man above 60 years and a child with blue eyes, are good customers*." is a classification rule which might be extremely significant in the examined database, but it is not probable that this rule can also be used for the classification of new customers. For this reason, a number of restrictions should often be imposed on the language in which the hypotheses are expressed.

B. *Support Vector Machines*

Classification and regression are probably the most-widespread applications of data mining. A multitude of techniques have therefore been proposed for solving these tasks. Linear least-squares regression, decision trees, discriminant analysis and neural networks are only a few of them. When confronted with a classification or regression problem, the data mining practitioner must often make a trade-off between the intelligibility and performance of the available techniques. For example, neural networks have proven to be excellent classifiers, but due to their complexity it is tough to understand why certain classification decisions are made. Classifications by decision trees on the other are clearly motivated by a number of rules that are represented by the tree. Techniques that are able to give the motivation behind

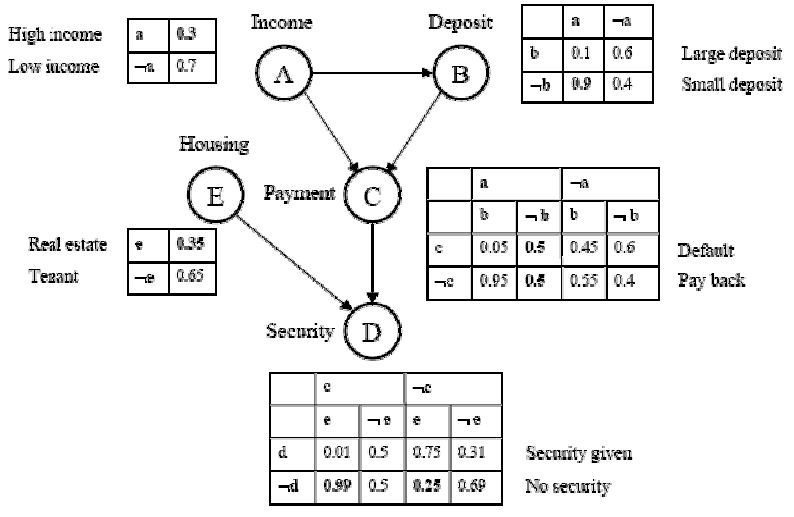
their decisions are called white-box classifiers. In recent years, lots of research has been performed to convert results from the difficult to understand black-boxes into white-boxes. For example, various rule extraction algorithms have been proposed for obtaining simple rules from a trained neural network trying to keep as much of the performance of the network as possible. Recently, a new black-box technique has been proposed that shows even better performance: support vector machines. The basic idea behind SVMs is the following: the data is first being mapped into a high-dimensional space and afterwards a linear classifier is constructed in this high- dimensional space. The resulting models can be represented as constrained optimization problems which give a unique solution.

C. *Bayesian Networks*

Besides support vector machines, we observe an increasing application of Bayesian networks within industrial applications. A Bayesian network is a graphical model where variables are presented by nodes and the edges between two nodes represent the dependencies between the variables.

Figure 3 shows a simple Bayesian network that can be used to decide on loan applications. Next to each node, a table indicates the conditional probabilities

FIGURE 3
Bayesian Network for credit scoring (Egmont-Peterson et al., 2004)



On the basis of this network one can make forecasts concerning the credit worthiness of a new applicant. For example, a person with a high income (a), a small deposit on his account (¬b), no security (¬d) and owner of his house (e) will get no loan of the bank because his probability of non-repayment exceeds the probability of repayment. These probabilities are calculated as follows:

$$P(c|a,\neg b,\neg d,e)=\frac{P(c,a,\neg b,\neg d,e)}{P(a,\neg b,\neg d,e)}=0{,}798$$

and

$$P(\neg c|a,\neg b,\neg d,e)=\frac{P(\neg c,a,\neg b,\neg d,e)}{P(a,\neg b,\neg d,e)}=0{,}202$$

with

$$P(c,a,\neg b,\neg d,e)=p(a)\,p(\neg b|a)\,p(c|a,\neg b)\,p(\neg d|c,e)\,p(e)=0{,}3*0{,}9*0{,}5*0{,}99*0{,}35=0{,}0467$$

$$P(\neg c,a,\neg b,\neg d,e)=p(a)\,p(\neg b|a)\,p(\neg c|a,\neg b)\,p(\neg d|\neg c,e)\,p(e)=0{,}3*0{,}9*0{,}5*0{,}25*0{,}35=0{,}0118$$

$$P(a,\neg b,\neg d,e)=P(c,a,\neg b,\neg d,e)+P(\neg c,a,\neg b,\neg d,e)=0{,}0467+0{,}0118=0{,}0585$$

Bayesian networks offer therefore a simple and clear way to take decisions under uncertainty. A disadvantage is however that one must first infer the structure of the network and provide values for the conditional probability tables. This can be performed with the aid of domain experts, but lots of research is done to use data mining techniques to derive the structure and the values from past data.

D. *Ensemble Methods*

The central idea behind ensemble methods is very simple: several classifiers are trained on the data and subsequently these individual forecasts are combined to obtain one general forecast. We will illustrate this technique with a small example. Suppose that for the next world championship football, a system must be developed that can predict the nationality of supporters (Belgian, Dutch, German). The first classifier bases itself mainly on the language and is able to recognize German fans but cannot distinguish between Belgian and Dutch supporters. The second classifier bases itself on the colors in which supporters are dressed and can recognize the orange of the Dutch supporters, but sees no distinction between the flags of Belgian and German supporters. If both classifiers are combined, we obtain a classifier that makes better forecasts than each of the individual classifiers separately.

There are several alternatives of ensemble methods: bagging, boosting and stacking.

With bagging, the abbreviation of bootstrap aggregating, n random subsets from the original data are selected and a classifier is created for each of these subsets. A new observation is then classified by combining the forecasts of the n classifiers.

Boosting shows similarities with bagging, but instead of selecting entirely random subsets, weights are given to the training observations. Observations that are misclassified receive a larger weight and their chance of being incorporated in the subset will increase because of this. Boosting therefore gives more attention to those observations which are more difficult to predict.

The third alternative, stacking, corresponds to the example of the World championship football. Several classifiers are trained on the available data and their forecasts are used as inputs for a so-called meta-learner. This meta-learner uses these individual forecasts to obtain one overall forecast.

The large disadvantage of these combination techniques is that they lead to a strong fall of intelligibility: it becomes very difficulty to explain why the model takes a certain decision.

V. CONCLUSION

It is impossible to imagine our society today without data mining. Both in scientific and industrial world, the applications have become too widespread. In this article, a short overview was given of some new domains in which data mining can cause immense changes. However, there are still many problems to overcome, from which privacy protection draws most attention. Privacy protection deserves certainly a solid amount of attention, but it should not lead to an exaggerated apprehension of data mining. After all, the possibilities and opportunities of data mining are too valuable, for example in the development cycle of new medicines.

In the second part of this article, some new techniques were evaluated. These techniques are still subject of further research, but we expect that they will make rapidly the transition into a business environment.

REFERENCES

CASPIAN, http://www.spychips.com/metro/scandal-payback.html.

Dublin Core Metadata Initiative, http://dublincore.org/.

Dzeroski, S., 1996, Inductive Logic Programming and Knowledge Discovery in Databases. Advances in Knowledge Discovery and Data Mining, (AAAI Press/The MIT Press), 117-152.

Dzeroski S., 2003, Multi-Relational Data Mining: an Introduction, *SIGKDD Explorations*, 5, 1, 1-16.

Egmont-Petersen M., Feelders A., Baesens B., 2004, Probabilistic Network Classifiers and Probability Confidence Intervals Illustrated by an Application in Credit Scoring, *Computational Statistics and Data Analysis*.

Fayyad U., Piatetsky-Shapiro G., Smyth. P., 1996, Knowledge Discovery in Databases: Towards a Unifying Framework, Second International Conference on Knowledge Discovery and Data Mining.

FOAF, http://www.foaf-project.org/.
Google News, http://news.google.com.
Huysmans, J., Baesens, B., Mues, C., Vanthienen, J., 2004, Web Usage Mining with Time Constrained Association Rules, Sixth International Conference on Enterprise Information Systems (ICEIS 2004), (Porto, Portugal).
Turing, A.M., 1950, Computing Machinery and Intelligence, (Oxford University Press, *Journal of the Mind Association*, 59, 236, 433-60.
Transport for London fact sheet, http://www.tfl.gov.uk/tfl/